Advanced Information Systems Technology

Empowering Data Management, Diagnosis, and Visualization of Cloud-Resolving Models by Cloud Library upon Spark and Hadoop

Completed Technology Project (2015 - 2017)

Project Introduction

A cloud-resolving model (CRM) is an atmospheric numerical model that can resolve clouds and cloud systems at very high spatial resolution. The main advantage of the CRM is that it can allow explicit interactive processes between microphysics, radia-tion, turbulence, surface, and aerosols. CRMs have played critical roles in many NASA satellite missions (TRMM, GPM, CloudSat) and science projects (MAP). Be-cause of their fine resolution and complex physical processes, it is challenging for the CRM community to i) visualize/inter-compare CRM simulations, ii) diagnose key processes for cloud-precipitation formation and intensity, and iii) evaluate against NASA's field campaign data and L1/L2 satellite data products due to large data vol-ume (~10TB) and complexity of CRM's physical processes. Rapid progress in com-puting technology (massive parallel computing, GPU) has exacerbated these chal-lenges by allowing larger-domain and higher-resolution CRM simulations without adequate support in data management. Objectives: The effects of aerosols on weather and climate are the largest uncertainty in predict-ing anthropogenic impact on the current weather and climate models. In this project, technology on Hadoop and Spark is used to empower database management, diagno-sis, and visualization of CRMs, and thus significantly improve the understanding of simulated processes associated with cloud-precipitation and their interaction with aerosols on weather prediction and climate change studies. Technical Status/Approach To this end, we propose to develop the Super Cloud Library (SCL), capable of CRM database management (IO control and compression), distribution, visualization, subsetting, and evaluation. SCL architecture is built upon a Hadoop framework. The Hadoop distributed file system (HDFS) is a stable, distributed, scalable and port-able file-system. The Hadoop framework supports Python, which enables 2D and 3D visualization through wrapping IDL codes. Further, Hadoop R enables various stan-dard/non-standard statistics and their visualization. Within the Hadoop framework, CRM's diagnostic capability will be further enhanced with Spark, built on top of HDSF, which accelerates Hadoop MapReduce process by ~100 times. SCL will be built on the NCCS Discover system, which directly stores various CRM simulations, including the NASA-Unified Weather Research and Forecasting (NU-Forecast (WRF) and Goddard Cumulus Ensemble (GCE) models. Thus, SCL users can conduct large-scale on-demand tasks automatically, without downloading volu-minous CRM datasets and various observations from NASA Field Campaigns and Satellite data to a local computer. This task will have a performance period of two years. During this time, we plan to take the Technology Readiness Level from an entry level of 2 (concept) to an exit level of 5 (system prototype in an operational setting).

Empowering Data Management, Diagnosis, and Visualization of Cloud-Resolving Models by Cloud Library upon Spark and Hadoop

Table of Contents

Organizational Responsibility

Responsible Mission Directorate:

Science Mission Directorate (SMD)

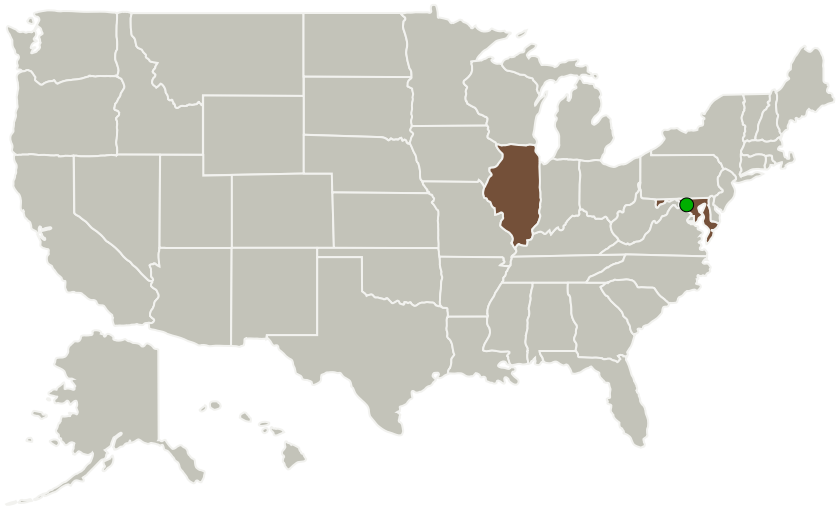
Responsible Program:

Advanced Information Systems Technology

Empowering Data Management, Diagnosis, and Visualization of Cloud-Resolving Models by Cloud Library upon Spark and Hadoop

Completed Technology Project (2015 - 2017)

Primary U.S. Work Locations and Key Partners



Organizations Performing Work	Role	Type	Location
Goddard Space Flight Center(GSFC)	Supporting Organization	NASA Center	Greenbelt, Maryland

Primary U.S. Work Locations	
Illinois	Maryland

Project Management

Program Director:

Pamela S Millar

Program Manager:

Jacqueline J Le Moigne

Principal Investigator:

Wei-kuo Tao

Co-Investigators:

Shujia Zhou
David T Leisawitz
Xiaowen Li
Xian-he Sun
Daniel Q Duffy
Toshihisa Matsui

Technology Areas

Primary:

- TX11 Software, Modeling, Simulation, and Information Processing
  - TX11.4 Information Processing
    - TX11.4.3 Semantic Technologies

Target Destination

Earth